2008 年度 修士論文

Web 掲示板上の評判情報における評価対象の仕様や一部分を表す表現の抽出

～複合語をグループ化することによる効率的な抽出手法～

早稲田大学大学院理工学研究科

情報・ネットワーク専攻

臼渕 護

学籍番号：3606U119 - 9

提出：2008 年 7 月 28 日

指導：山名 早人 教授

概要

近年インターネットの普及に伴い，掲示板やブログ等のテキストデータが大量に Web 上に蓄積されるようになった．これらのテキストデータから製品に対する個人の評価が記述された評判情報を抽出することにより，開発を行う企業はもちろん，購入を考えている消費者にも有益な情報をもたらすことができる．評判情報において評価を受けるのは製品の仕様（性質や特徴）や一部分を表す属性表現であるため，評判情報抽出を行うためにはこれらの辞書を「プリンタ」や「車」などの製品のドメイン毎にあらかじめ作成しておく必要がある．この際，属性表現の多くは二つ以上の単語が組み合わさってできた複合語であり，これらをどう扱うかで抽出の仕方は異なってくる．

従来手法では例えばプリンタドメインにおける「高画質モード」や「フォトモード」のように属性表現が複合語である場合，その一段上の概念である上位語「モード」として大まかに扱う手法とそのまま複合語として個別に扱う手法の 2 種類が提案されている．前者のメリットとして上位語のみを扱えば良いため辞書に登録すべき表現が少なくて済む点が挙げられるが，「高画質モード」と「フォトモード」などの上位語が等しい複合語同士の区別ができないことに加え，造語などの語句としての妥当性が低い複合語や「i モード」や「留守電モード」などのプリンタドメインに関連性のない複合語も属性表現として扱ってしまうデメリットがある．それに対し後者の手法の場合，より精密に評判情報を扱えることに加え，高精度の評判情報抽出が可能になるが，大量の表現を登録する必要があるため精度，再現率の高い辞書を構築するためには多大なコストがかかるというデメリットがある．

本論文ではこういった問題に対し，複合語を上位語とせずそのまま扱いながら属性表現の辞書作成を行った場合でも低コストで精度，再現率の高い辞書を作成する手法を提案する．提案する手法では，前述のプリンタドメインにおける「高画質モード」と「フォトモード」のように同じドメイン内で上位語が等しい複合語を同一視して，属性表現の抽出を行う際に紛れ込んでしまう造語やドメインに関連性のない語句などのノイズをあらかじめフィルタリングする．こうすることで，同一ドメインで上位語が等しい複合語をひとまとめにして扱うことが可能になり，複合語の属性表現を一つ一つ別々の語句として扱いながら辞書作成を行う従来手法に比べ効率的な辞書作成が可能になった．

実験の結果，プリンタドメインでは精度 86%，再現率 91%，バイクドメインでは精度 89%，再現率 93%で複合語の属性表現の抽出がされ，本手法の有効性を示すことができた．

目次

1 はじめに

近年 Web 上に大量に蓄積された掲示板やブログなどのテキストデータから，評判情報を収集する技術[1][2][3][4]に注目が集まっている．評判情報とは，「印刷が汚い」や「美しい画像だ」のような，ある製品に対する個人の評価のことであり，製品の開発を行っている企業はもちろん，購入を考えている一般の人々にとっても有益な情報である．これらの評判情報において評価を受けるのは製品の仕様（性質や特徴）や一部分を表す属性表現である．そのため，評判情報を抽出する際にはあらかじめ製品のドメイン毎に属性表現辞書を作っておくことが非常に重要になってくる．

ここで，属性表現の多くは「フォトモード」「高画質モード」といったような複合語である．評判情報抽出の再現率を向上させるためには，複合語の属性表現を抽出することが必須である．複合語の属性表現をどのように抽出するかという点で，従来手法の抽出アプローチは次に示す 2 つに分けられる．属性表現が「高画質モード」や「フォトモード」のように複合語からなる場合を考える．1 つ目のアプローチは，「モード」という上位語として大まかに扱うアプローチであり，2 つ目のアプローチはそのまま複合語として個別に扱うアプローチである．

前者のアプローチとして小林ら[1]は，属性表現が「きれいだ」や「美しい」のように属性表現を評価する表現（以下評価表現と呼ぶ）と何種類かのパタンを形成して出現する傾向があることに注目し，いくつかの種となる評価表現をもとに半手動でブートストラップ的に上位語の属性表現と評価表現の辞書を作成する手法を考案した．また峠ら[2]は，小林らの手法[1]が属性表現の候補を辞書に登録するかどうかの最終的な判断に人手を要しているのに対し，これらの判断を自動で行えるように改善処理を加えることで辞書構築の効率化を行った．これらの上位語の属性表現のみを対象とした辞書構築[1][2]の場合，登録すべき表現が少ないため精度よく網羅的に収集することは容易である．しかし，例えばプリンタドメインにおいて「フォトモード」と「高画質モード」の区別ができないことに加え，「i モード」や「留守電モード」などのドメインに関連性のない語句を属性表現として扱ってしまうなど，精密さに欠ける．

これに対し，後者のアプローチとして，小林ら[3]は自らが提案した手法[1]に属性表現を複合語で扱えるような処理を加えることで複合語の属性表現を半手動で辞書登録することを可能にした．また峠ら[4]は「属性表現はメインクエリ（「プリンタ」や「車」などの製品名）の周辺に存在している」という考えに基づき，メインクエリとのテキストデータ内での距離を利用して上位語に加え複合語の属性表現も自動で抽出できる手法を提案した．これらの複合語の属性表現も抽出可能な手法[3][4]によって作成された辞書の場合，より詳細に評判情報を扱うことができるが，辞書作成時に多くの表現を辞書登録する必要があり，高精度で再現率の高い辞書を作成することは困難である．

本論文ではこうした問題に対して，複合語のまま属性表現を扱った場合でも低コストで

精度よくかつ網羅的に辞書構築を行う手法を提案する．複合語の属性表現の辞書構築をする際，前述のプリンタドメインにおける「フォトモード」と「高画質モード」のようにドメインが同じでかつ上位語が等しい語句の場合，意味的に近いため，同一のものとして扱ったほうが効率的である．しかし一方で上位語が等しくても，造語やスペルミスのように語句としての妥当性が低い表現や前述の「i モード」や「留守電モード」のようにプリンタドメインに関連性のない語句，形態素解析の誤りによってできた語句などのノイズとなる複合語も多数存在する．そこで本提案手法ではまずドメインに存在するすべての複合語に関して前述のノイズとなる語句をフィルタリングする処理を行う．フィルタリングを残った複合語は語句としての妥当性があり，かつドメインに関連性があるので上位語が等しければ同一語句として扱いながら辞書の作成を行う．

本論文では，以下の構成をとる．第 2 章において関連研究について説明し，第 3 章で評判情報における用語の定義をする．続いて第 4 章において提案手法の説明を行い，第 5 章において提案手法の実験，及び考察を行う．そして最後に第 6 章においてまとめを行う．

2　用語定義

本章では本論文で使われる用語に関しての定義を行う．まず評判情報を構成する要素に関して定義を行い，その後属性表現の抽出単位に関して定義を行う．

2.1　評判情報の構成要素の定義

評判情報抽出の研究をまとめた奥村らの調査[5]によると，評判情報の構成要素は以下のように定義されている．

ⅰ．対象表現

「プリンタ」や「BMW」のように評価の対象となる商品名、及びサービス名のことを指す．

ⅱ．属性表現

対象表現の仕様（性質や特徴）や一部分などのことを指す．例えばプリンタドメインにおいては「画質」や「インク」，自動車ドメインにおいては「ハンドル」や「ブレーキ」などが属性表現になる．

ⅲ．評価表現

属性表現の質や量に対しての評価を表す表現である．一般的な評価表現として「美しい」「便利だ」「汚い」「不便だ」などが挙げられる．

2.2　属性表現の抽出単位の定義

辞書作成の際，属性表現の抽出単位は以下の二つに分かれる．

ⅰ．複合語

「フォトモード」や「デジタル画像」のように二つ以上の単語が組み合わさってできた語句．

ⅱ．上位語

「フォトモード」に対する「モード」や「デジタル画像」に対する「画像」のようにある複合語が属するグループを表す語句．複合語の構成単語の最後尾にある一単語からなる．

3　関連研究

本章では属性表現抽出の関連研究について述べる．これらは属性表現抽出の際，複合語を上位語として扱うアプローチと，複合語のまま扱うアプローチの 2 通りのアプローチがある．それぞれのアプローチについて，該当する関連研究を述べる．

3.1　複合語をその上位語とみなした場合の属性表現の辞書構築手法

複合語をその上位語とみなして辞書作成を行う手法[1][2]について説明する．本手法では「デジタル画像」のような複合語はその上位語である「画像」として扱っている．

3.1.1　小林ら[1]の手法

2003 年に奈良先端科学技術大学院大学の小林ら[1]は評判情報を＜対象，属性，評価＞の 3 要素からなると定義し，人手を介しながら上位語の属性表現と評価表現の辞書を作成する手法を考案した．この手法はブートストラップに基づいており，以下のような共起パタンを介して，評価表現と属性表現を相互に獲得する．

共起パタン：＜属性＞｛が/は/を/も/に｝＜評価＞
例）画質はきれいだ

例えば上記の共起パタンに当てはまる例の場合，「きれい」が評価表現であることがわかっていれば「画質」のスコアがプラス 1 される．小林らは以下に示す 8 種類の共起パタンを用意しており，これらを利用してテキストデータからスコアが一定値以上の語句を属性表現の候補として抽出する．なお，全ドメイン共通の評価表現辞書をあらかじめ作成し用意している．

パタン 1. ＜評価＞＜対象＞
　例）便利なプリンタ
パタン 2. ＜評価＞＜属性＞
　例）美しい画像
パタン 3. ＜評価＞＜属性＞
　例）かっこいいデザイン
パタン 4. ＜対象＞の＜属性＞
　例）プリンタの画質
パタン 5. ＜属性＞｛が/は/を/も/に｝＜評価＞
　例）リモコンが使いやすい
パタン 6. ＜属性＞｛が/は/を/も/に｝＜評価＞
　例）効果音が小さい
パタン 7. ＜対象＞の＜属性＞｛が/は/を/も/に｝＜評価＞

例）BMW の座席は心地よい

パタン 8. ＜対象＞の＜属性＞｛が/は/を/も/に｝＜評価＞

例）プリンタの印刷速度が遅い

（下線部は候補表現として抽出される部分を表す）

次に候補の中から妥当性の高い表現を人手で判定し，属性表現として辞書登録する．そして新たな表現が登録された辞書を利用して，今度は評価表現と属性表現の抽出を行う．

以後，これらの処理を繰り返し行うことでブートストラップ的に辞書の作成を行う．実験の結果，コンピュータの商品名を含む Web 文章約 9 万文に対し 201 個の上位語の属性表現が抽出されている．

3.1.2 峠ら[2]の手法

3.1.1 の小林らの手法では，属性表現かどうかを決定する為に手作業が必要になってしまう欠点がある．したがって，ドメインごとに辞書作成をしたい場合にはかなりのコストがかかってしまう．そこで 2004 年に長岡技術科学大学の峠ら[2]は上位語の属性表現をドメインごとに自動で抽出する手法を提案した．抽出は次の 2 ステップを踏む．

ステップ 1. 属性表現の候補を抽出

小林ら[1]が作成した共起パタンと独自に作成した評価表現の辞書を利用してスコアリングを行い，スコアが 1 以上の語句を属性表現の候補として抽出する．

ステップ 2. ノイズの自動フィルタリング

ステップ 1 で抽出した属性表現の候補に含まれるノイズは大まかに分けて 2 種類があると考え，それぞれに以下のようなフィルタリングを行う．

① スコアが低い属性表現の候補をフィルタリング

属性表現の候補の中には，スコアが低いものも多かったが，それらのほとんどは属性表現として不適当な語句であった．そこでスコアが一定値以下の語句を属性表現の候補から削除する．

② 一般的な語句のフィルタリング

属性表現の候補の中には「ハンドル」や「ブレーキ」などのドメインに特化した語句のほかに，「人」「月」「年」等の一般性の高い語句が多く含まれている．

しかしこれらの語句の多くは製品の仕様や一部分を表してはおらず，属性表現として不適当であった．そこで新聞や Web コーパスを利用し，これらのテキストデータ中で高頻度で出現する一般性の高い語句を属性表現の候補から削除する．

これら 2 ステップの後，残った語句を属性表現として登録する．ただし自動で辞書作成ができる反面，登録すべき属性表現を削除したり，ノイズをうまく削除できていなかったりするミスもあり，低頻度ノイズ，高頻度ノイズ共にフィルタリングの精度，再現率は 50

～60%となっている．

3.2　複合語をそのまま扱った場合の属性表現の辞書構築手法

複合語をそのまま扱いながら属性表現の辞書作成を行う手法[3][4]について説明する．本手法では「フォトモード」と「高画質モード」のように上位語が等しい複合語が同一ドメイン内にあっても別々の語句として扱う．

3.2.1　小林ら[3]の手法

小林らは自らが考案した 3.1.1 の手法を，複合語の属性表現も辞書登録できるように拡張した．こうすることで「デジタル画像」と「アナログ画像」を別々に扱うことができ，詳細な評判情報抽出が可能になった．しかし 3.1.1 の手法と同様に属性表現かどうかの最終的な判断は人手で行っているため辞書構築に手間がかかってしまう．特に複合語の属性表現の数は膨大であり，ドメインによっては多大なコストがかかる．「車」と「ゲーム」のレビューサイトから収集した記事を対象にして実験を行った結果，抽出された表現数と要した時間の関係を以下の表にまとめる．

表 1，小林らが収集した属性表現の数と要した時間（[3]より引用）

ドメイン名	記事数	抽出に要した時間（人手（一名）で抽出した時間も含む）	抽出した表現数
車	230,000 文	7 時間	3,965 個
ゲーム	90,000 文	5 時間	2,631 個

抽出された表現の数はドメインにより異なるが，大体 2,000～4,000 個であり，これらを半手動で抽出するために要する時間は 5～7 時間となっている．そのためこれらの手法を利用して辞書構築を行うには多くの人手と時間が必要になってくる．

3.2.2　峠ら[4]の手法

3.2.1 で小林らが提案した手法は十分な人手が確保できる場合は有効な手法である．しかしそうでない場合，ドメインごとに辞書作成を行うことは困難である．そこで峠らは複合語を含めた属性表現を自動で抽出する手法を提案した．峠らは「属性表現はメインクエリ（「車」，「プリンタ」などの製品名）に対して関連の深い語と考えることができるため，テキストデータのおいて，メインクエリの周辺に多く存在している」という仮定の下，以下の 5 ステップを踏むことで抽出を行った．なお，文中で現れる隣接語，連想語が本研究での属性表現に当たる．

ステップ 1．候補となる語句を同定

テキストデータ中に存在する候補となる語句の同定を行う．候補となる語句は「名詞－一般」,「名詞－サ変接続」,「名詞－固有名詞」,「未知語」と複合語である．複合語の構成要素には「名詞－一般」,「名詞－サ変接続」,「名詞－固有名詞」,「未知語」,「記号列」を利用している．

ステップ 2．一文単位で語句のペアを作成

隣接する候補のペアを一文単位で抽出する．

例）この**車**の**エンジン**にもう少し**トルク**があれば**運転**も楽しくなるのに．

→｛車，エンジン｝,｛エンジン，トルク｝,｛トルク，運転｝

ステップ 3．メインクエリを利用した隣接語（属性表現）の抽出

メインクエリの周辺には属性表現が現れやすいという仮定の下，メインクエリとペアになっている語句を隣接語（属性表現）として抽出する．

例）｛**車**，エンジン｝→「エンジン」を隣接語とする．

｛加速，**車**｝→「加速」を隣接語とする．

ステップ 4．連想語候補の抽出

隣接語とペアになる語句をステップ 3 のメインクエリの連想語候補として抽出する．

例）エンジンが隣接語の場合

｛**エンジン**，ブレーキ｝→「ブレーキ」を連想語候補（前検索）

｛ハイブリッド，**エンジン**｝→「ハイブリッド」を連想語候補（後検索）

ステップ 5．連想語の抽出

ある隣接語の連想語候補が前検索と後検索の両方に含まれていた場合，その語句を連想語として抽出する．

例）エンジンが隣接語の場合

前検索：「**ブレーキ**，**エンジンオイル**，水，**ハイブリッド**，ミラー，…」

後検索：「**ブレーキ**，**エンジンオイル**，角度，水平方向，**ハイブリッド**，上がり，…」

→　ブレーキ，エンジンオイル，ハイブリッドが連想語となる．

以上のステップで得られた隣接語，連想語を属性表現として辞書登録している．価格.com[]掲示板から収集した「携帯電話」,「車」,「デジタルカメラ」の書き込みに対して実験を行った結果を以下の表に示す．

表 2，峠らが収集した属性表現の数とその精度（[4]より引用）

ドメイン名	記事数	抽出した表現数	精度
携帯電話	850000 文	3503 個	71%
車	1060000 文	7122 個	80%
デジタルカメラ	1160000 文	5803 個	76%

自動で辞書構築ができるため大量のテキストデータを対象にして抽出ができ，多くの属性表現を獲得できる．しかし抽出された属性表現の精度は 70～80%となっており，ノイズが多い．

3.3 関連研究のまとめ

これまで述べてきたように属性表現の辞書を作成する際，複合語を上位語として扱うか，そのまま扱うかで 2 通りのアプローチが提案されている．

前者のメリットとして「デジタル画像」と「アナログ画像」のような上位語が等しい表現を個別に抽出する必要がなく，「画像」としてひとまとまりにくくれるため，抽出すべき表現数が少なくて済む点があげられる．しかしその反面，「デジタル画像」のような複合語は辞書登録できないため，より具体的に評判情報を扱いたい場合には不向きである．

また後者のメリットとして「デジタル画像」や「アナログ画像」のように上位語が等しい語句であっても別々に扱えるため，より詳細に評判情報を扱えるという点があげられる．しかしその反面，辞書に多くの表現を登録する必要があり，精度，再現率の高い辞書を作成するためには多大なコストがかかってしまう．

そこで本論文では複合語をそのまま扱いながら属性表現の辞書作成を行った場合でも低コストで精度，網羅性の高い辞書を作成する手法を提案する．

表 3，関連研究のまとめ

	複合語の扱い	辞書作成のコスト
複合語を上位語として扱う手法[1][2]	不可	低
複合語をそのまま扱う手法[3][4]	可	大
提案手法	可	低

4　提案手法

本章では属性表現の抽出システムについて説明する．属性表現を抽出する際には，複合語をその上位語としておおまかに扱う手法[1][2]より，個別に扱う手法[3][4]のほうがより精密な評判情報抽出が行える．しかし，従来手法[3][4]では複合語の属性表現を抽出する際に，「デジタル画像」と「アナログ画像」のように上位語が等しい複合語を別々に扱いながら辞書作成を行っているため，抽出すべき属性表現の数が多く，精度，再現率の高い辞書を作成するためには多大なコストがかかってしまっていた．

そこで本提案手法では上位語が等しい複合語をグループ化して扱うことで低コストで精度，再現率の高い辞書作成を行う．ただし，単純に上位語が等しい複合語をグループ化すると造語やドメインに関連性のない語句などのノイズが紛れ込んでしまうため，まずこれらのノイズを自動でフィルタリングし，その後グループ化を行いながら属性表現の抽出を行う．

提案するシステムは**＜1＞テキストデータからの複合語の抽出**，**＜2＞ノイズのフィルタリング**，**＜3＞属性表現の抽出**，の 3 つのモジュールからなる．全体の構成を図 1 に示す．

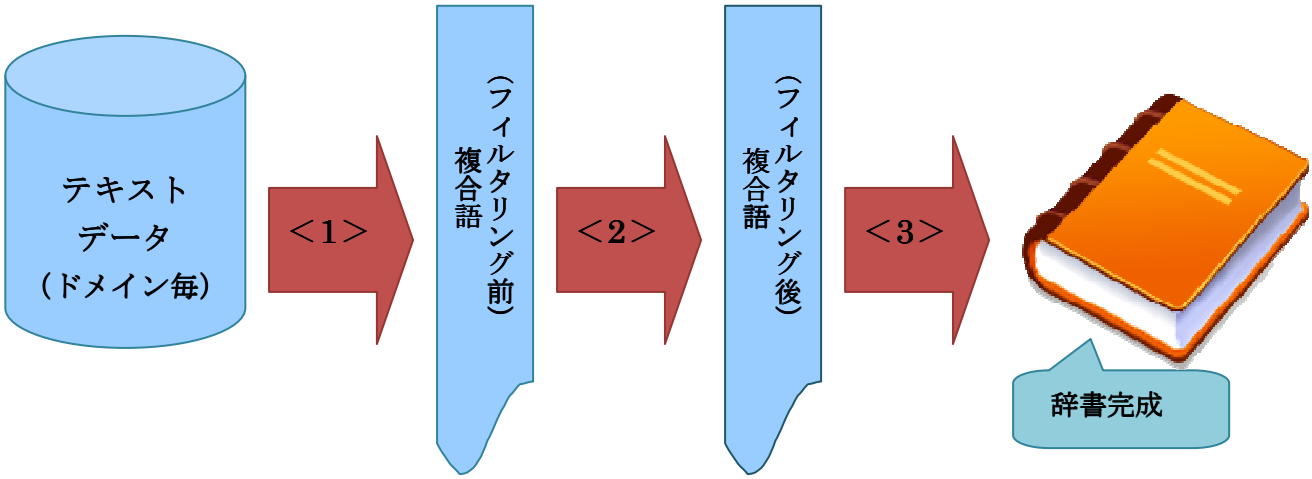

図 1，属性表現の抽出システム

4.1　テキストデータからの複合語の抽出

ドメイン毎に収集したテキストデータを南瓜[6]で形態素解析し，複合語の抽出を行う．この際，複合語は形態素解析を行うことにより，例えば「カラー印刷」は「カラー（名詞－一般）」＋「印刷（名詞－サ変接続）」のように別の形態素に分割されるため，複合語としてまとめあげる必要がある．構成要素は品詞体系で「名詞－一般」，「名詞－サ変接続」，「名詞－固有名詞－一般」，「記号－アルファベット」と解析された語句とする．なお，複合語の構成要素の最後尾の語句，つまり上位語に当たる語句は品詞体系で「名詞－一般」，

「名詞－サ変接続」と解析された語句のみとする．

4.2 ノイズのフィルタリング

4.1 で抽出した複合語から事前にフィルタリングしたいノイズとは，以下の 3 種類に分類される複合語のことである．具体例としてそれぞれのノイズについてプリンタドメインにおいて上位語が「モード」である場合を示す．

1. 造語やスペルミス等の妥当性の低い語句

例）自動電源 ON モード，両面モード，拡張解像度モード

2. 形態素解析の誤りにより生じた語句

例）から印刷モード，最初高画質モード

3. ドメインに関連性のない語句

例）i モード，留守電モード，美白モード

これらのノイズをフィルタリングすることができれば，同一ドメインで上位語が等しい語句は語句としての妥当性もあり，意味的にも近いと考えられるのでひとまとめにして扱うことが可能になる．

そこでこれらのノイズについてフィルタリング機能を設ける．4.2.1～4.2.3 でそれぞれのノイズのフィルタリング手法について説明する．

4.2.1 造語やスペルミス等の妥当性の低い語句のフィルタリング

個人が勝手に作った造語やスペルミスのように妥当性の低い語句をフィルタリングする．これらの語句の特徴として，使用する人がほとんどおらず，Web 上に存在するサイト内での出現頻度が少ないことがあげられる．そこで Yahoo! Japan[7]のフレーズ検索のヒット数 h_w を利用して，これらのノイズのフィルタリングを行う．

ただし，フレーズ検索では検索エンジンが以下に示されるような記号や空白を接続記号として認識している．

－（ハイフン）　/（スラッシュ）　.（句読点）　＝（等号）　　など

そのため例えば「方法印刷」や「染料顔料」という造語をフレーズ検索した際に，それぞれ「…方法（画質）…」や「…染料 顔料…」といった文が含まれるサイト数もカウントされ，h_w が本来の語句の出現サイト数より多く算出されてしまう．

そこで複合語をクエリとした検索結果の上位 100 件のスニペットのうち，接続記号や空白を含まない状態でその複合語が出現する数 P_w を求め，それを利用して h_w を以下のように正規化する．

$$h_w' = h_w \times \frac{P_w}{100} \qquad (1)$$

その後h_w'が閾値以下の語句をノイズとしてフィルタリングする．ただし，ある語句が妥当な語句であるかどうかを判断する際，人によって誤差が出てきてしまうので今回は峠ら[4]の研究を参考にして閾値を設定する．峠らはある表現が語句として妥当かどうかの判断にYahoo! Japan[7]での検索ヒット数を利用しており，その閾値を1,000と設定している．そこで本研究でもこれにならい閾値を 1,000 と設定し，次式に当てはまる語句をフィルタリングする．

$$h_w' \leq 1{,}000 \tag{2}$$

4.2.2 形態素解析の誤りにより生じた語句のフィルタリング

「副詞－助詞類接続」や「助詞－格助詞」などの 4.1 で複合語の構成要素として指定しなかった語句が誤って形態素解析され，「名詞－一般」などの指定した品詞として判定されてしまった結果生じた語句をフィルタリングする．これらの解析誤りが起こってしまう語句について調べたところ，「あまり」や「あと」のように使われ方によっては複合語の構成要素として指定しなかった語句（この場合は「副詞－助詞類接続」）とも，「名詞－一般」とも解析される可能性がある語句であることがわかった．使用した形態素解析器[6]では，例えば動詞の後に格助詞が来ることはない，といったような規則をもとに品詞の解析を行っているので，日本語として不自然な書き込みがあった場合，これらのミスは避けられない．ただし，これらの日本語として不自然な書き込みというのは Web 上に存在する割合としてはそれほど多くないと考えられる．

そこで試しにこれらのノイズをクエリとした検索結果のスニペットをいくつか抜き出し形態素解析を行った．その結果，ほとんどのスニペットで解析誤りが起こった品詞に関して，正しく形態素解析が行われた．例えば「あまり」が「副詞－助詞類接続」ではなくて「名詞－一般」と解析された結果生じた「あまり印刷」というノイズの場合，「あまり印刷」をクエリとした検索結果のスニペットに形態素解析を行ったところ，ほとんどのスニペットで「あまり」が「副詞－助詞類接続」として正しく解析されていた．このことから形態素解析誤りが起こってしまうような日本語として不自然な書き込みというのは，Web 上に存在する割合としては非常に少ないことがわかった．

そこで検索結果のスニペットを利用してこれらのノイズを自動でフィルタリングする処理を設ける．具体的には以下の 4 ステップを踏むことでフィルタリングを行う．

1. **複合語wをクエリとした検索結果の上位 100 件のスニペットを取得する**
2. **取得したスニペットに形態素解析を行う**
3. **wを構成する品詞が 4.1 で指定した構成要素のみであるスニペットの数Q_wを数える**
4. **Q_wが次式で示される閾値T_aを下回る複合語をノイズとしてフィルタリングする**

$$Q_w \leq T_a \tag{3}$$

4.2.3　ドメインに関連性のない語句のフィルタリング

語句としての妥当性は高く，形態素解析の誤りなどもないものの，ドメインに対しての関連性がないため，上位語が等しい他の複合語と同一視することはできないと判断される語句のフィルタリングを行う．

まず，これらの語句をフィルタリングするためにドメイン関連度D_wを定義する．ドメイン関連度D_wとは，その語句が Web 上に存在するサイト内でどれだけドメイン名（「プリンタ」，「車」などの製品名）と多く共起しているかを利用して関連度の強さを求めた指標であり，次式で表わされる．

$$D_w = a \times \frac{h_{w,d}}{h_w \times h_d} \tag{4}$$

a　：正規化係数

h_d：ドメイン名dをクエリとした Yahoo！ Japan[7]でのフレーズ検索のヒット数

$h_{w,d}$：複合語wとドメイン名dをそれぞれダブルクォーテーションでくくり，Yahoo！ Japan[7]で AND 検索した場合のヒット数

ドメインに対する関連性が低い語句の特徴として，このドメイン関連度D_wが低いことがあげられる．そこで次式で示されるドメイン関連度Dが閾値T_dを下回る語句をフィルタリングする．

$$D_w \le T_d \tag{5}$$

4.3　属性表現の抽出

4.2 でフィルタリングを行った複合語から属性表現を抽出する．抽出は以下の 3 ステップを踏むことで実現される．

ステップ 1．文型パタンと評価表現の辞書によるスコアリング

以下に示す小林ら[1]の提案した文型パタンと人手で作成した評価表現の辞書とを利用して，複合語にスコアリングを行う．なお，評価表現の辞書は現代形容詞用法辞典[8]を参考にし，使用頻度が高くかつ一般性がある語句を独自の判断で 250 表現用意した．

1. ＜属性表現＞が/は/も/に/を　［評価表現］
   例）〈カラーインク〉が［汚い］
2. ［評価表現］　＜属性表現＞
   例）［美しい］＜印刷画質＞

例えば，「接続スピードが遅い」という書き込みがあったとする．今，「遅い」が評価表現辞書に登録されているとすると，上記 1 に合致し，「接続スピード」のスコアに 1 が加算される．

ステップ 2. 上位語が等しい複合語のグループ化

「フォトモード」と「高画質モード」のように上位語が等しい複合語をその上位語である「モード」のグループとして統一する．その際，グループ u のスコア S_u は u を上位語とする複合語の集合 C_u に含まれる複合語 w のスコア $score_w$ を用いて(4)式で表わされる．

$$S_u = \sum_{w \in C_u} score_w \tag{6}$$

例えば，「モード」を上位語とする複合語が「高画質モード」と「フォトモード」のみであり，前者がスコア 3 で，後者がスコア 6 であれば，「モード」グループのスコアは 9 になる．

ステップ 3. 属性表現のグループの抽出

ステップ 2 で作成されたグループに関して，次式で示されるスコア S_u が閾値以上のグループを抽出する．

$$S_u \geq T_g \tag{7}$$

この際，閾値を高く設定すると精度よく属性表現のグループを抽出できる反面，網羅性に欠けてしまう．また閾値を低く設定すると精度よく抽出はできないものの，網羅的な抽出が可能になる．本手法ではできる限り多くの属性表現を精度よく得るために閾値を低めに設定して抽出を行い，そこから手作業で属性表現のグループか否か分類を行う．

分類の際は，そのグループに属する複合語が属性表現としてふさわしいかどうかを判断基準とする．正しいと分類されたグループの複合語は属性表現として辞書登録する．

5　実験と考察

5.1　テキストデータの収集と複合語の抽出

提案手法の評価対象として，2008 年 6 月時点で価格.com[9]のプリンタ及びバイクのカテゴリに書き込まれていたレビューを対象とした．まず，プリンタ，及びバイクのレビューを収集し，これらをテキストデータとして複合語を抽出した．書き込みの総数と抽出された複合語の異なり数を表 3 に示す．

表 3，ドメイン毎に収集した書き込みの総数と抽出された複合語の異なり数

ドメイン名	書き込みの総数	複合語の異なり数
プリンタ	108,956 文	19,362 個
バイク	325,196 文	13,382 個

5.2　フィルタリングの閾値設定，および考察

フィルタリングの閾値を設定し，どの程度精度良く，かつ網羅的にノイズを削除できているのか，またフィルタリングの誤りがなぜ起こってしまうのかついて考察を行う．閾値設定の際は，それぞれのフィルタリングに関して閾値を変化させながら(8)式のように定義される *F*値を求め，値が最も高くなるところに閾値を定める．*F*値は適合率と再現率の調和平均であり，フィルタリングがどれだけ精度よく網羅的にノイズを削除できているのかを示す．

$$F\text{値} = \frac{N_{fn}}{\frac{1}{2}(N_{fw} + N_n)}$$

N_{fn} :フィルタリングされたノイズの数
N_{fw} :フィルタリングされた複合語の数
N_n :実際にノイズである複合語の数
（ノイズは著者が手作業でカウントした）　　(8)

なお 4.2.1 で前述したように，造語やスペルミス等の妥当性の低い語句のフィルタリングに関しては判断が人によって異なってしまう可能性があるため閾値はあらかじめ一定値に定め，フィルタリングに対する考察のみを行う．

5.2.1　造語やスペルミス等の妥当性の低い語句フィルタリング

フィルタリングを行うことで，多くの造語やスペルミス等の妥当性の低い語句をフィリタリングすることができた．ただし，フィルタリングは h_w' が 1000 以下であれば無条件に行ったため，専門性が強く，Web 上であまり使用されることがない結果 h_w' が低くなってい

る語句などは妥当性があっても削除されてしまう結果となった．これらの語句についてドメイン毎の具体例を表 4 に示す．

表 4，h_w'が 1000 以下だが妥当性のある属性表現

プリンタ	ペーパーフィードカセット，CD レーベルコピー，モノクロ写真印刷，インククリーニング
バイク	ファーストマフラー，くるぶしグリップ，ミドルスクリーン，プレストモデル

h_w'が 1000 以下の語句を造語やスペルミス等の妥当性の低い語句としてフィルタリングした結果，フィルタリング前後で複合語の数がどの程度変化したかを表 5 に示す．

表 5，造語等の妥当性の低い語句のフィルタリングをした時の複合語の数の変化

ドメイン名	フィルタリング前	フィルタリング後
プリンタ	19,362 個	17,649 個
バイク	13,382 個	11,892 個

5.2.2　形態素解析の誤りにより生じた語句のフィルタリング

形態素解析誤りにより生じた語句のフィルタリングに関して閾値T_aを 5～95 まで 5 づつずらしながら前述のF値を求めた結果を図 2（プリンタ），図 3（バイク）に表わす．

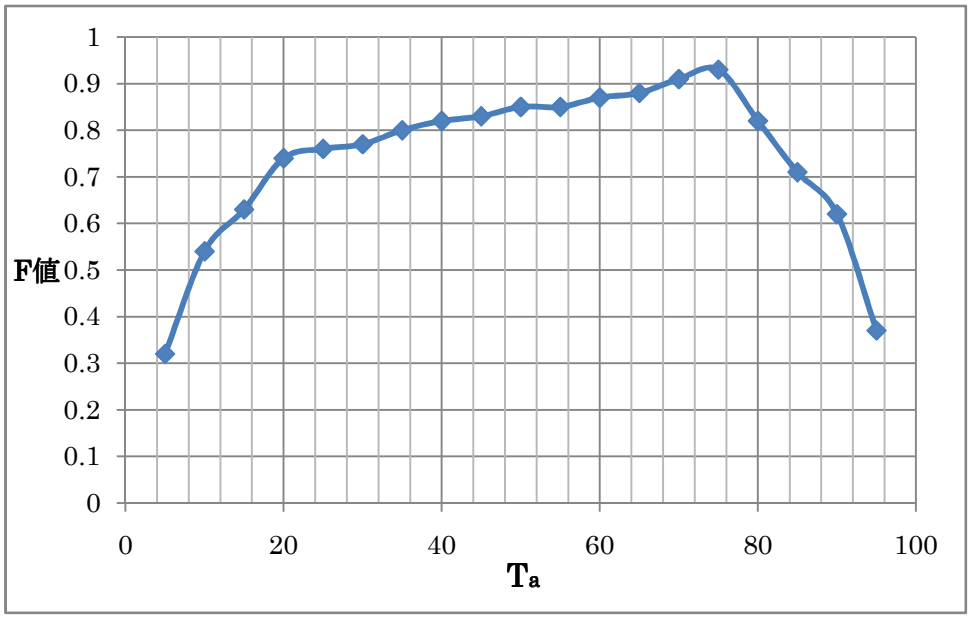

図 2，プリンタドメインにおける閾値T_aとF値の関係

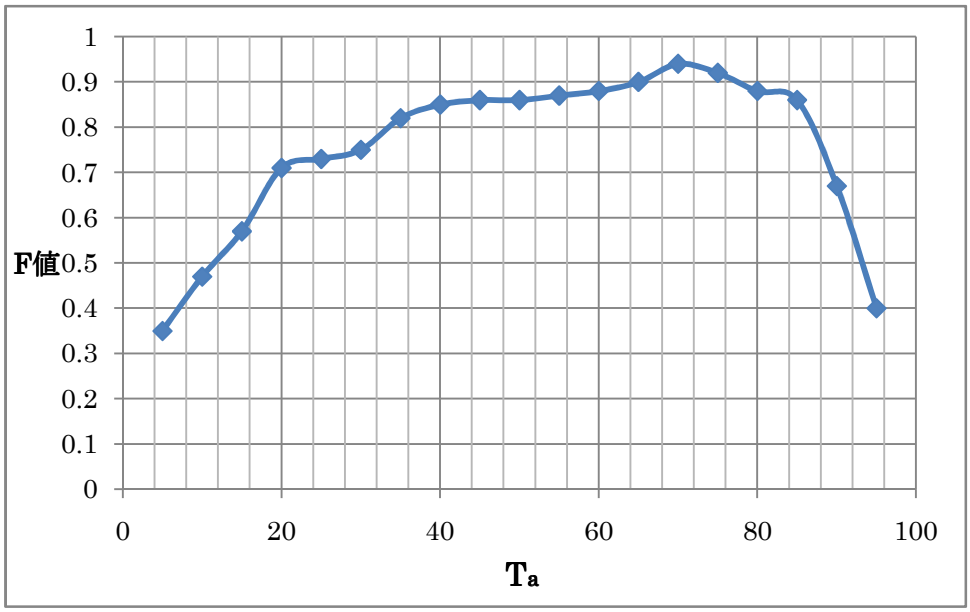

図 3，バイクドメインにおける閾値T_aとF値の関係

図 2 と図 3 から閾値T_aをプリンタドメインに関しては 75，バイクドメインに関しては 70 と定めればF値が最も高くなることがわかる．またこのときF値は 0.90～0.95 の値を示しており，精度よく網羅的にノイズを削除できていることがわかった．プリンタ，バイクの両ドメインとも閾値T_aが 65～75 の時にF値が 0.85 以上になっていることから，実験を行っていない他のドメインに関しても閾値T_aを 65～75 に定めた時にF値が最も高くなると考えられる．

フィルタリングの際に起こってしまうミスの原因を調べるため，　Q_wについて考察した．Q_wは複合語wをクエリとした検索結果の上位 100 件のスニペットのうち，wを構成する品詞が 4.1 で指定した構成要素のみであるスニペットの数である．結果，形態素解析の誤りにより生じた語句は多くが 0～40 という低い値を示したが，一部 40～90 の高い値を示したことがわかった．また誤りのない語句に関しては多くが 80 以上の高い値を示したが，一部 60～80 と低い値を示す語句もあることがわかった．これらのフィルタリングミスの要因になる語句の具体例を表 6 にまとめる．

表 6，フィルタリングミスの要因となる語句

ドメイン名	形態素解析誤りがあるのに Q_wが高い語句	形態素解析誤りがないのに Q_wが低い語句
プリンタ	程度印刷，最初インク	コピー画面，液晶解像度
バイク	数回ブレーキ，多分マフラー	リアタイヤ，ギアオイル

使用した形態素解析器[6]では「最初」や「程度」のような一部の単語を無条件で「名詞

- 一般」と解析してしまうため，これらは避けられない誤りであることがわかった．一方で誤りの部分が「から」や「あと」のようにひらがなである語句の多くはQ_wが低くカウントされ正確にフィルタリングされていた．

また，形態素解析の誤りがないのにQ_wが低い語句に関しては，検索結果のスニペットにその語句が出現しない場合が多く，Q_wを正しくカウントできていないことがわかった．検索エンジンではある語句をクエリとして検索を行った時にそれに合致する内容のサイトがあれば，たとえサイト内にその語句が出現しなくても検索結果に含めることがあるためだと考えられる．

閾値T_aをプリンタドメインに関しては75, バイクドメインに関しては70と定めたとき，フィルタリング前後で複合語の数がどの程度変化したかを表7に示す．

表7，形態素解析の誤りにより生じた語句のフィルタリングをした時の複合語の数の変化

ドメイン名	フィルタリング前	フィルタリング後
プリンタ	17,549 個	16,961 個
バイク	11,892 個	11,154 個

5.2.3　ドメインに関連性のない語句のフィルタリング

ドメインに関連性のない語句のフィルタリングに関して閾値T_dを1～12まで1づつずらしながら前述のF値を求めた結果を図4（プリンタ），図5（バイク）に表わす．なお，正規化係数aは1000000とした．

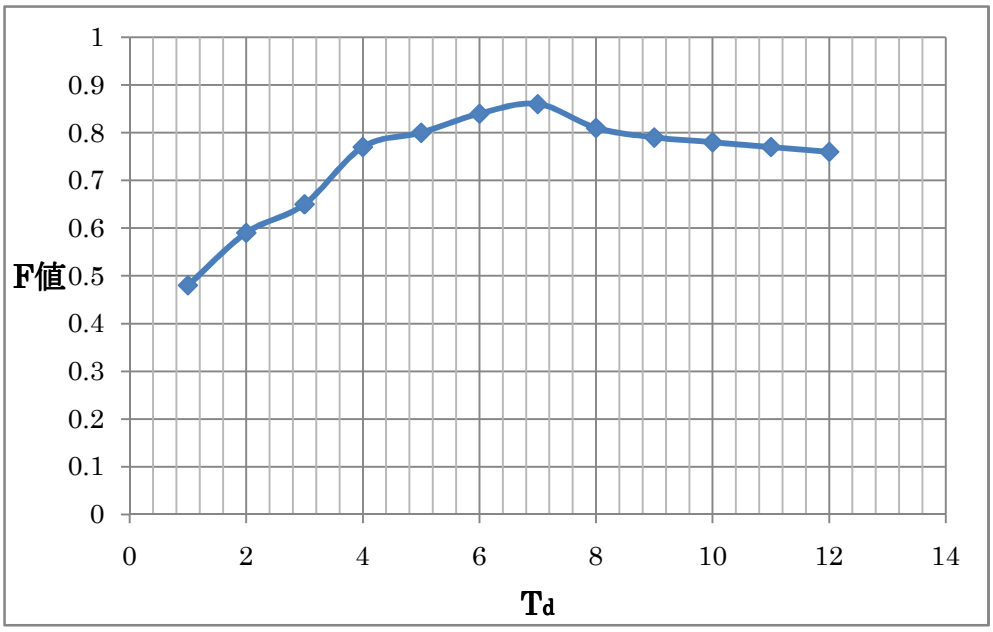

図4，プリンタドメインにおける閾値T_dとF値の関係

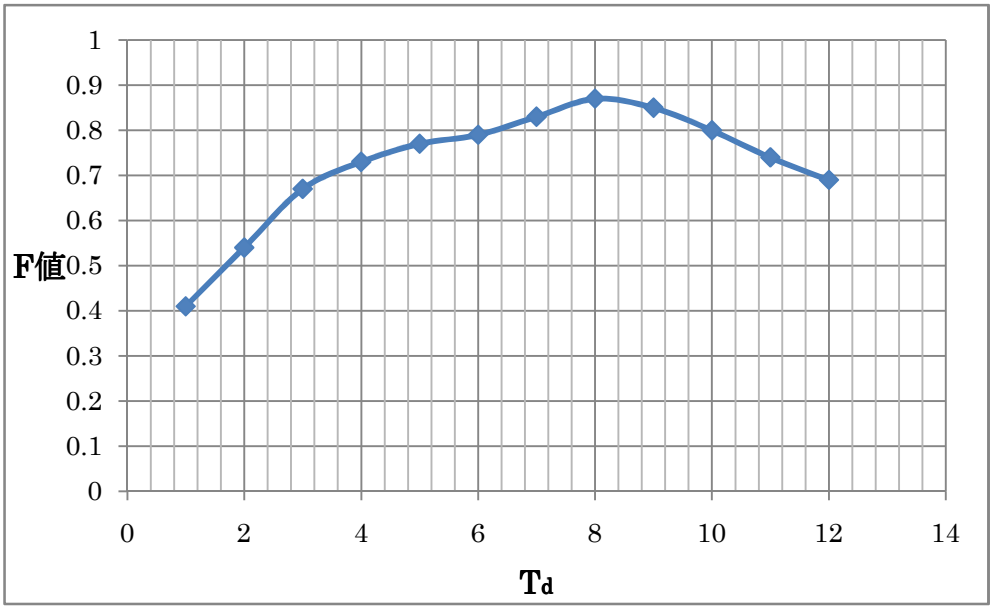

図 5，バイクドメインにおける閾値 T_d と F 値の関係

図 4 と図 5 から閾値をプリンタドメインに関しては 7，バイクドメインに関しては 8 と定めれば F 値が最も高くなることがわかる．またこのとき F 値は 0.8～0.9 の値を示しており，精度よく網羅的にノイズを削除できていることがわかった．プリンタ，バイクの両ドメインとも閾値 T_a が 7～9 の時に F 値が 0.8 以上になっていることから，実験を行っていない他のドメインに関しても閾値 T_a を 7～9 に定めた時に F 値が最も高くなると考えられる．

フィルタリングの際に起こってしまうミスの原因を調べるため， D_w について考察した．D_w は複合語 w が Web 上に存在するサイト内でどれだけドメイン名（「プリンタ」，「車」などの製品名）と多く共起しているかを利用してドメインに対する関連度の強さを求めた指標である．結果，ドメインに関連性のない語句の多くは D_w が低く算出されていたが，一部高く算出されてしまう語句もあることがわかった．またドメインに関連性がある語句の多くは D_w が高く算出されていたが，一部低く算出されてしまった語句もあり，これらがフィルタリング誤りの要因になったと考えられる．フィルタリングミスの要因となった語句の具体例を表 8 にまとめる．

表 8，フィルタリングミスの要因となる語句

ドメイン名	ドメインに関連性がないのに D_w が高い語句	ドメインに関連性があるのに D_w が低い語句
プリンタ	パソコン画面，撮影モード	CD トレイ，カラー写真
バイク	自転車タイヤ，ドアロック	クラッチオイル，バックライト

ドメインに関連性がないのに D_w が高い語句の多くは，プリンタに対するパソコンやカメ

ラ，バイクに対する自動車や自転車のようにそのドメインと関連性の強い他のドメインの属性表現であった．

またドメインに関連性があるのにD_wが低い語句の多くは，一般性が強く，h_wが高くカウントされた結果，D_wが低くなってしまっていた．

閾値T_dをプリンタドメインに関しては 7，バイクドメインに関しては 8 と定めたとき，フィルタリング前後で複合語の数がどの程度変化したかを表 9 に示す．

表 9，ドメインに関連性のない語句のフィルタリングをした時の複合語の数の変化

ドメイン名	フィルタリング前	フィルタリング後
プリンタ	16,961 個	11,827 個
バイク	11,154 個	7,318 個

5.3 属性表現の抽出結果

表で示される適切な値に閾値を設定し，属性表現の抽出を行った．それぞれのステップでの出力結果を表 10 に示す．

表 10，設定した閾値

ドメイン名	T_a	T_d	T_g
プリンタ	75	7	10
バイク	70	8	10

表 11，属性表現抽出の各ステップでの出力結果

ドメイン名	複合語の数（＊）	グループ数	S_uがT_g以上のグループ数	属性表現のグループ数	属性表現の数
プリンタ	11,827 個	1,517 個	356 個	191 個	3,215 個
バイク	7,318 個	1,036 個	312 個	159 個	1,921 個

（＊）フィルタリング後の複合語の数である．

プリンタドメインに関しては精度 86%，再現率 91%，バイクドメインに関しては精度 89%，再現率 93%で抽出が行われた．使用したテキストデータが異なるため，単純に他の研究と比較はできないが高精度，高再現率で抽出が行われたことがわかる．また属性表現のグループを抽出する際にかかった手作業の時間はそれぞれ 15～20 分であり，低コストで抽出が行われたことがわかる．

5.4 複合語をグループ化して扱うことによるメリット

複合語をグループ化して扱うことにより，どのようなメリットがあるのかを調べるため，

$score_w$ 毎の属性表現とノイズの数をプリンタドメインは図 6, バイクドメインは図 7 に示す.

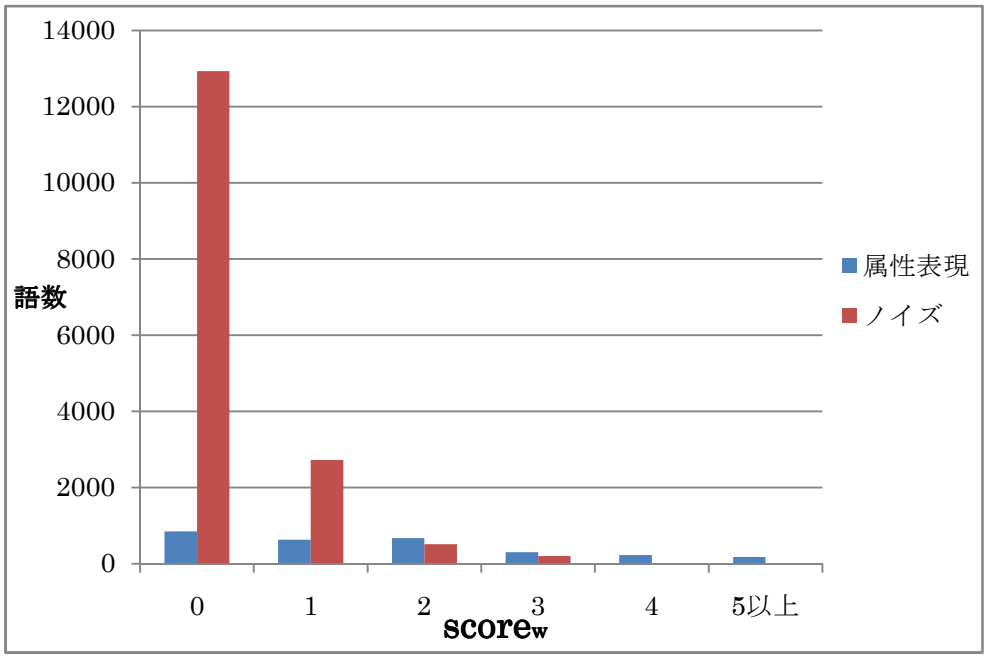

図 6, プリンタドメインにおける $score_w$ 毎の属性表現とノイズの数

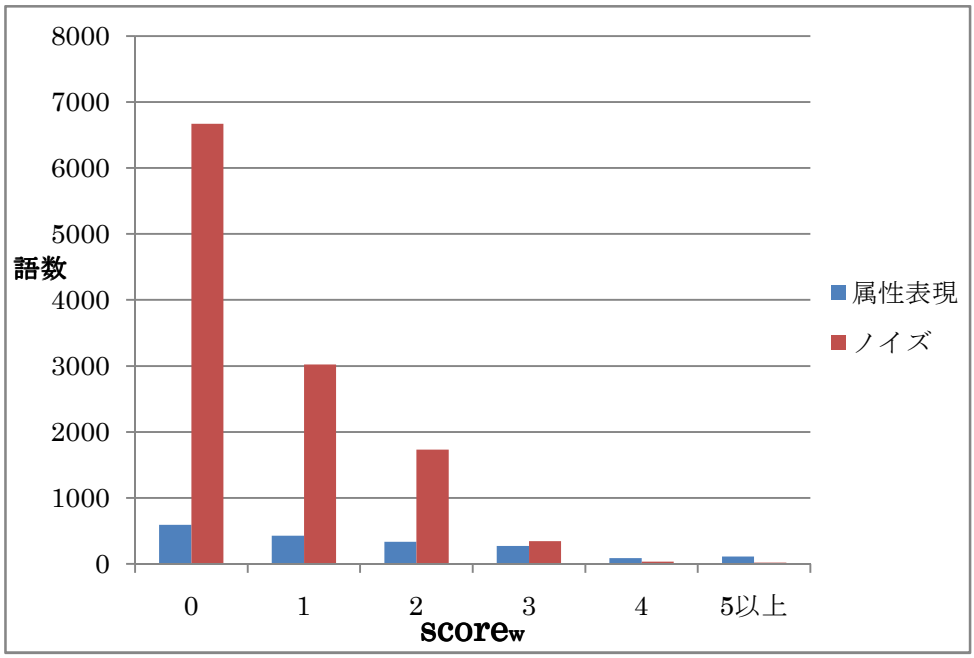

図 7, バイクドメインにおける $score_w$ 毎の属性表現とノイズの数

図 6, 図 7 より複合語の属性表現を一つ一つ別々のものとして扱った場合, 例えば $score_w$ が 1 以上の複合語を属性表現として抽出すると, 膨大な数のノイズも抽出してしまうこと

が分かる．そのため精度の高い抽出を行うためには閾値を 4 以上に設定しなくてはならなず，$score_w$ が 3 以下の語句は抽出が困難である．しかし，本論文で提案した手法によって複合語をグループ化して扱うことにより，$score_w$ が 3 以下の属性表現をプリンタドメインにおいては精度 84%，再現率 89%，またバイクドメインにおいては精度 86%，再現率 91%で抽出できる．このことから複合語をグループ化して属性表現抽出をおこなうことにより，$score_w$ が低い属性表現を精度よく網羅的に抽出できたことがわかる．

5.5 その他のノイズ

抽出の精度を下げる要因となった，フィルタリングを試みたノイズ以外のノイズについて考察する．

5.5.1 構成要素に指定しなかった語句により構成される複合語

「再利用インク」の「再」（接頭詞－名詞接続）や「CD－R メディア」の「－」（記号－一般）などの構成要素として選択しなかった品詞に該当する語句が含まれる複合語は，それぞれ「利用インク」や「Ｒメディア」のように意味を成さないノイズとして抽出されてしまう．そのためこれらを構成要素として追加する必要があるのだが，単純に「接頭詞-名詞接続」や「記号－一般」と解析される語句をすべて構成要素にしてしまうとノイズが急激に増えてしまうため，これらの品詞の中でも構成要素にして良い語句としてはいけない語句を分類する必要があると考えられる．表 12 に再検討が必要だと考えられる品詞をまとめる．

表 12，複合語の構成要素として再検討が必要な品詞

接頭詞－名詞接続，接頭詞－数接続，記号－一般，名詞－数， 固有名詞－人名，固有名詞－組織，固有名詞－地域，未知語

5.5.2 抽象的な複合語

ドメインに関連性はあり，形態素解析の誤りもないのだが，辞書に登録するには不適当な語句があった．具体的には「各社」，「当社」，「上記」などの抽象的な語句よって修飾される語句である．これらの語句は h_w'，D_w が共に高いのでフィルタリングを通過してしまうのだが，辞書に登録しても具体性がなく利用価値がない語句であった．

6 まとめ

本論文で提案した手法を用いることにより得られた成果を以下にまとめる.

- グループ化を行うことで一つ一つ別々の複合語として扱う従来手法[]に比べ抽出コストを大幅に減らすことができた.

- 複合語の属性表現に関して，プリンタドメインで精度 86%，再現率 91%，バイクドメインで精度 89%，再現率 93%という高精度，高再現率で抽出が行われた.

- グループ化を行うことで $score_w$ が低い語句に関しても高精度で網羅的な抽出が可能になった.

また，今後改善していく必要がある部分に関して以下にまとめる.

- h_w' が低いためフィルタリングされてしまうものの妥当性は高い語句がいくつかあった. そこで今後は収集したテキストデータ内での出現頻度なども考慮し，この値が大きい語句は h_w' が 1000 以下であってもフィルタリングしないようにするなどの改善が必要である.

- 形態素解析の誤りがないのに Q_w が低く，フィルタリングされてしまう語句がいくつかあった．検索結果のスニペットにその語句が出現しない場合があり，Q_w を正しくカウントできていないためである．そのため今後は検索で使用した語句が含まれていないスニペットは利用しないようにするなどの工夫が必要である.

- ドメインに関連性があるのに D_w が低く，フィルタリングされてしまう語句がいくつかあった．これに対して以下の二通りの改善策が考えられる.
  - 「“ドメイン名”の“複合語”」でフレーズ検索してそのヒット数が一定値以上の語句はフィルタリングしないようにする．例えば表の例の場合「プリンタの CD トレイ」と検索した時のヒット数が一定値を超えた場合,「CD トレイ」はフィルタリングしないようにする.
  - 収集したテキストデータ内での出現頻度が多い語句はそれだけドメインへの関連度が強いと考えられるため，フィルタリングしないようにする.

謝辞

参考文献

[1]小林 のぞみ, 乾 健太郎, 松本 裕治, 立石 健二, 福島 俊一(2003). “テキストマイニングによる評価表現の収集” 情報処理学会研究報告, NL154-12 pp77-84.
[2]峠 泰成, 山本 和英(2004). “手がかり語自動取得による Web 掲示板からの評価文抽出” 言語処理学会第 10 回年次大会, pp107-110.
[3]小林 のぞみ, 乾 健太郎, 松本 裕治, 立石 健二, 福島 俊一(2005). “意見抽出のための評価表現の収集” 自然言語処理, 12(2), pp203-222.
[4]峠 泰成, 山本 和英(2006). “意見情報獲得のためのクエリー関連のドメイン特徴語抽出” 言語処理学会第 12 回年次大会, pp85-88.
[5]乾 孝司, 奥村 学(2006). “テキストを対象とした評価情報の分析に関する研究動向” 自然言語処理, Vol13, Num3, pp201-241.
[6]日本語係り受け解析器“南瓜”. http://chasen.org/~taku/software/cabocha/
[7]Yahoo! JAPAN. http://www.yahoo.co.jp/
[8]飛田良文, 浅田秀子(1991). “現代形容詞用法辞典” 東京堂
[9]価格.com 掲示板. http://kakaku.com/bbs